EFFICIENT DATA GENERATION FOR MACHINE LEARNING: INTEGRATING DESIGN OF EXPERIMENT AND DATA QUALITY

Ann-Kathrin Koschlik*, Lars Steffens†, Florian Raddatz*, Gerko Wende*

* Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Instandhaltung und Modifikation, Hein-Saß-Weg 22, 21129 Hamburg

† Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für KI-Sicherheit, Rathausallee 12, 53757 Sankt Augustin

Abstract

A good detection algorithm and its underlying model rely fundamentally on a high-quality data foundation. However, the immediate question arises: What constitutes a high-quality data set? While large volumes of operational data can comprehensively reflect realistic system, newly designed systems lack access to such data during early life cycle phases. Consequently, data synthesis during the design phase becomes increasingly important, particularly when preparing experiments for subsequent machine learning and deep learning applications. In the context of machine health systems, experimental data acquisition is often costly, time-consuming, or infeasible due to safety risks and system constraints. Therefore, the number of physical experiments must be minimized, while ensuring that the generated data remains informative and representative for downstream modelling tasks. This paper lays the groundwork for data synthesis and optimal data set design, leveraging structured and statistical methodologies. The discipline of Design of Experiments (DoE) has been well established in academic and industrial domains as an efficient approach to systematically explore input–output relationships under constrained experimental budgets. By integrating DoE principles with simulation-based data generation and machine learning requirements, the proposed approach enables the creation of informative, diverse, and statistically robust data sets, even in the absence of extensive real-world observations. The concept of data quality is well established and standardized. Data quality follows the principle of fitness for use and ISO/IEC 25012 defines data quality as "degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions". Many methods are available for assessing the quality of data sets which are being integrated. This method is demonstrated for the use-case of acoustic fault detection of unmanned aerial vehicle propellers.

Keywords

Quality Engineering, Design of Experiments; Data Quality; Reliability; Causality; Experimental design

Contact address:

ann-kathrin.koschlik@dlr.de

AI Artifical Intelligence
AL Assurance Levels
BPF Blade Pass Frequency
BW Beam Width
DoE Design of Experiments
DR Dynamic Range
DQ Data Quality
DQMo Data Quality Model
DQMa Data Quality Management
DQR Data Quality Requirements
CBM Condition-Based Maintenance
CNN Convolutional Neural Network
ConOps Concept of Operations
FHA Functional Hazard Assessment
FQ Frequency
$\mathbf{L}_p$ Sound Pressure Level
ML Machine Learning
OD Operational Domain
ODD Operational Design Domain
PdM Predictive Maintenance
PHM Prognostic and Health Management
PSF Point Spread Function
$\mathbf{r}_{psf}$ Pearson Correlation of Point Spread Function
$\mathbf{r}_{AIM}$ Pearson Correlation of Acoustic Image
SPL Sound Pressure Level
TQL Tool Qualification Levels
UAV Unmanned Aerial Vehicle
UAS Unmanned Aerial System

1. INTRODUCTION

Successful adoption of data-driven diagnostics in industry hinges on several obstacles: (1) names the reliability and accuracy of Predictive Maintenance (PdM) algorithms and limited Data Quality (DQ) as central, while noting that better data collection and quality control drive up algorithm cost. According to EASA Regulation (EU) 2024/168, the trustworthiness of Artifical In-

telligence (AI) is a fundamental prerequisite for the use of AI in aviation, supported by the building blocks of AI assurance, human factors, and risk mitigation. EUROCAE ED-324 / SAE ARP6983 provides a recognized means of demonstration for supervised learning Machine Learning (ML) components, provided that all assurance level objectives are met and substantiated by the required lifecycle data. Central to this is end-to-end traceability between the AI system's operational design domain and the operational domain, including requirements for Data Quality, preprocessing, and feature engineering, as well as the handling of nominal, edge, corner, and outlier cases in the risk process. Against this background, its regulatory implications around the handling and generation of data are of great importance, which gives rise to the need for methodological support in the design of AI systems. The discipline of DQ which is commonly defined by the the principle of "fitness for use" (2; 3), states that "data are of high quality if they are fit for their intended uses by customers" (4). Several studies underline the urge for a systematic need of DQ methodologies: (5) situate DQ within the Prognostic and Health Management (PHM) context and show that quality deficiencies do not become apparent until late in the process, which is why they introduce detectability as a metric against data quality requirements; (6) identify uncertainty, robustness as essential properties of predictive methods and propose an evaluation framework.

Beyond test-set accuracy, trustworthy ML emphasises the reliability of predictions on new, potentially shifted data. In the context of AI-based system risk assessment, uncertainty must be made explicit (7): epistemic uncertainty, arising from limited knowledge or insufficient training data, and aleatory uncertainty, arising from inherent noise and randomness. The former is typically assessed via deep ensembles, Bayesian models, Monte Carlo dropout, or out-of-distribution detection (8); the latter through predictive variance, noise models, confidence intervals. Their combination yields the total predictive uncertainty, indicating the reliability and confidence of model predictions (7). ML is particularly fruitful where the relationship between input and output data is unknown, the analytical modelling equations are highly complex or nonlinear (9, S. 1841). Yet, the high demand in data may result in high experimental costs. Consequently, the data acquisition strategy becomes highly relevant at an early point of time: Design of Experiments (DoE) is a long-established approach to data collection in which the factor combinations to be run are laid down in advance as a single experimental design, allowing the relevant variables to be identified and the response optimized within a constrained experimental budget (10). A reciprocal benefit exists between DoE and ML is the identifying the ideal algorithmic parameter settings, DoE can render ML algorithms more efficient, while ML in turn substantially supports the central aim of DoE by identifying the causality relationships (9, S. 1845). In recent years, this research direction has gained more interests as shown by systematic reviews of the interplay between DoE and ML are provided by (11; 12). The studies confirm, that the minimization of the number of experiments is achieved when combining DoE and ML. Simultaneously, the challenges of ML application due to its sensitivity to DQ as primary cause. PHM includes fault detection as a primary step. Multiple modelling approaches exist, mainly ranging from data-driven models to physics-based models. Their capability and performance strongly depend on the quality, structure, and representativeness of the available data. With the increasing relevance of proactive maintenance strategies, the implementation, validation, and certification of such models become highly important.

The data quality model proposed by (5) investigates the impact of data quality dimensions on fault detection performance, extending beyond commonly considered industrial dimensions such as data volume, data accuracy, and data completeness. By relating detectability to dimensions such as imbalance ratio, missing data, and noisy data, the authors derive an empirical data quality model for PHM applications. However, their work mainly evaluates data quality from a retrospective perspective, i.e., after data have already been acquired. Consequently, questions concerning how to design new data acquisition strategies, develop suitable tools, and plan experiments to obtain high-quality data sets at an early design stage remain unanswered.

This use-case study shifts the perspective from retrospective data quality assessment towards prospective data quality planning. In particular, it investigates how the systematic design of experiments can support the generation of complete, representative, and high-quality data sets for PHM applications. Based on the assumption that the quality of PHM models depends not only on the amount of data available, but also on how these data are generated, structured, and distributed, the following research question is formulated:

How can Design of Experiments and Data Quality be used at an early design stage to systematically generate high-quality data sets that improve the reliability and performance of PHM fault detection models?

2. METHODOLOGY

2.1. Data Quality

ISO/IEC 25012 defines DQ as the "degree to which the characteristics of data satisfy stated and implied needs under specified conditions", and lists those characteristics, most importantly accuracy, completeness, consistency, credibility, currentness, and compliance. (13) A frequent approach for assessing data quality involves establishing a Data Quality Model (DQMo), a "defined set of characteristics which provides a framework for specifying data quality requirements and evaluating data quality," as defined in the ISO/IEC 25012 standard (13). In this context, characteristics — more commonly referred to in the literature as quality dimensions (14) — are defined as "categories of data quality attributes that bears on data quality". The term "characteristics" can be applied to quantitative and qualitative assessment, whereas the term "dimension" more specifically refers to a "measurable feature of data" (14). An illustrative list of char-

acteristics for creating a DQMo can be found, for example, in ISO/IEC 25012, in ISO/IEC 5259-2, the ISO standard for DQ for ML (15), and in (3). An overview of numerous characteristics and their selection can be found in (14; 16). For instance completeness and accuracy are characteristics used in almost every DQMo. Besides this qualitative description a DQMo needs "measures" or "metrics" to quantify the quality, with a proper scale allowing proper assessment conclusions (17). Exemplary lists of metrics for special cases can be found in (15; 18) and, for sensor data, in ISO 8000, Parts 210 and 220, (19; 20).

As mentioned, (5) introduces a simple DQMo that was developed specifically for PHM and consists of three characteristics — completeness, accuracy, and volume, for each of which two metrics are introduced to evaluate the quality of the entire dataset. In this publication, in which we aim for certifiability, we focus on DQMo based on the relevant standards and expand the concept to include relevant aspects from the literature. Besides the already mentioned ISO standards, for our use-case this extends to the ISO standard for recording the noise of drones (21) and compliance standards for aeronautical data (22).

2.1.1. Process of Data Quality Modelling

This section proposes an iterative process for creating suitable DQMos. This process makes it possible to describe and define the relevant aspects of DQ for each part of the system and its analysis, independently of the other methods in our framework. It comprises the following steps:

Step 1) Identify the relevant data entities for the given use-case. The term "data entity" refers here to any specific level of the data hierarchy, such as a specific data item, a specific data record, or the entire dataset. In the literature, the dataset is often considered as a whole for the sake of simplicity. Here, we are taking the detailed approach, as we are aiming for certification and admission of the entire system. As illustrated below for various data entities within the dataset, the same characteristic can have different meanings across the different data entities, and also generally requires different metrics.

Step 2) Identify the requirements for the different data entities by first identifying relevant characteristics for each relevant data entity. To do this, characteristic catalogues from both standards and literature can be used as a guide. The application of current standards supports the certification process, while consulting the latest technical literature enables the use of state-of-the-art information. In this publication we use the standards (13; 15) the framework from (3) and the reviewing articles (14; 16). In our experience, adapting the definitions of the generally precisely formulated standards to the specific use-case—without distorting the meaning of the standard—enables the identification of potential quality issues in the system and paves the way for establishing appropriate metrics.

Step 3) Identify the characteristics that may be relevant or linked to DoE. Also identify characteristics that are primarily determined or influenced by the system design, so that the system can be designed from the ground up to produce "good data".

Step 4) Implement the metrics of the DQMo. For high amounts of identified data entities, characteristics and metrics, this represents a significant effort, that must be taken into account and traded-off with the general effort to design the system.

Step 5) Conducting data pollution experiments. In this step, the effects of "bad data" on the system are investigated by deliberately polluting the data, and the impact on parameters of other system aspects—such as the evaluations of the ML model—is examined. Basic studies investigating these effects of simple DQ characteristics on ML methods were conducted by (23). Two types of pollution can be distinguished: experimental pollution, where the experiment itself is manipulated to represent non-ideal circumstances—for example, by introducing physical noise—and synthetic pollution, where high-quality experimental data is artificially polluted afterwards. Although the first type of pollution more accurately reflects reality, it makes the experimental campaigns significantly more complex, whereas the second type is easier to apply to the data but may lack accuracy. A crucial aspect of the DoE in this context is finding a balance between accurate experimental data and the effort required to conduct the experiments. The data pollution studies allow to determine what effects certain types of non-ideal data have on the system output; in essence, they examine the system's robustness.

A step of data optimization is not considered in this framework, but might be necessary for specific data entities.

2.2. Design of Experiment

Variance in measurement results and noise variables are often inevitable. The observed variances across the measured data result in reduced accuracy. This systematic error is caused often wrong adjustment such as incorrect calibration of the apparatus or wrong handling, etc. Suitable experiments avoid systematic error. Individual wrong measurement may results in random errors, and their appearance is randomly. They impact the precision of the measurement data and their cause stems from experimental error. (24)

The purpose of statistical experimental design, or commonly referred to as DoE, is the detailed plan for data gathering by means of experiments. Classical main objective is the analysis of functional relationship. Industrial companies such as Bosch apply DoE successfully across their product creation process since 1988. (24) Whereas at an early design phase the cause-effect relationship is unknown, the process approach assumes that deeper knowledge is gained along the completion of the product. As the result, the DoE sharpens the required and missing knowledge from an initial cause-effect relationship for new product, to functional relationship between predictor and response variable up for improved product design to a robust model parametrization for various settings during the operations. This highlights the universal approach of DoE.

A DoE distinguishes system variables (describing the system and its operating conditions), prediction variables or factors (the deliberately varied inputs), and response variables (the measured outputs to be explained). The design is developed in sequence: fixing the responses and factors, the factor range and model approach, the factor levels and treatments, the number of replications, the run order and grouping, and finally the test procedure, equipment, specimens, and an estimate of time and effort.

2.3. Linking Data Quality to the Experimental Design: What Constitutes a High-Quality Dataset?

This sub-section proposes a framework that makes the DQ requirements explicit for the PHM context, by separating three levels at which a dataset can be assessed.

Inherent level: Are the recorded values correct, are records complete, is the dataset internally consistent, and can each sample be traced back to the run that produced it. None of this depends on the downstream task, and all of it can be checked before a model is trained. It is the entry condition for everything that follows.

Experimental design level: The second level concerns the experimental plan rather than the recorded numbers, and this is where DoE contributes. Relevant here are the coverage of the factor space, the alias structure of the design, and the statistical power available at the planned number of runs — properties of the plan that are fixed before the first measurement is taken.

Task level: The third level asks whether the dataset actually supports the learning task. Coverage can be stated as the fraction of the operational design domain that the plan occupies, representativeness as a divergence between the training distribution and the one expected in operation, and sufficiency as the saturation of the learning curve.

The two disciplines act at opposite ends of the same process. DoE is prescriptive: it specifies which factor combinations are to be run and at which levels, and the choice of design is itself iterative (Figure 1). DQ is descriptive: its measures quantify what the executed campaign actually produced, and so determine whether the plan achieved what it was designed for. DoE makes no statement about calibration errors, channel dropout or clipping; these are properties of the measurement chain and fall entirely to DQ. In reverse, DQ has no concept of experimental architecture. A dataset can satisfy every quality characteristic and still be useless because it covers a single operating point, and an optimal design can yield meaningless numbers if the acquisition chain is not capable.
The first two levels are model-independent and can be verified before a single measurement is taken; the third becomes visible only once a model has been trained. We therefore do not argue that DoE produces high-quality datasets in an absolute sense. It maximizes the expected information per run, provided the factor space that was modelled contains the space encountered in operation — and that is an assumption, not a result. What DoE does provide is the only means of influencing data quality *before* acquisition; everything else in the quality discussion acts afterwards. This also opens a further step: once the deviations recorded by DQ are known, they need not be treated only as defects to be detected, but can be introduced deliberately as additional factors, extending the range of parameter values covered by the design.

2.4. Use-Case: Acoustic Emission Fault Detection of Unmanned Aerial Vehicle (UAV) Propeller Blades and Key Features

2.4.1. Fault Detection System

The proposed framework is demonstrated on the acoustic fault detection of UAV propeller blades (25). The case is well suited to the methodology for three reasons. Propeller condition is safety-critical, as a degraded blade directly affects thrust, stability, and airworthiness. The relationship between a defect and its acoustic signature is complex and hard to capture analytically, while labelled fault data for newly designed systems is scarce — precisely the conditions that motivate a data-synthesis and design-of-experiments approach. Finally, the microphone array yields spatially rich acoustic maps in which faults manifest as changes in the localised source pattern, making the case a demanding but instructive testbed whose factor effects can be examined both in aggregate and across the array geometry.
The diagnostic value of the acoustic emission lies in the fact that different fault mechanisms leave distinct traces in the signal. A healthy propeller produces a characteristic signature dominated by the blade-passing frequency and its harmonics, superimposed on a broadband aerodynamic component. Faults modify this signature in mechanism-specific ways, as summarised in Table 1: a motor or rotor imbalance shifts the blade frequency, mechanical issues introduce unusual harmonic patterns, and aerodynamic problems manifest as increased broadband noise. Each fault mode thus maps to an identifiable change in the acoustic signature — in the fundamental frequency, the harmonic pattern, or the broadband level — forming the physical basis for the diagnostic task.

TAB 1. Key Features and Fault Mechanisms

Fault Mechanism	Key Feature
Motor/Rotor imbalance	BPF
Mechanical issues	Unusual harmonic pattern
Aerodynamic problems	Increased broadband noise

The system combines several technologies across the sensing, processing, and classification chain. At the front end sits an acoustic camera — the CAE Bionic M112, a 112-microphone MEMS array arranged in a circular, seven-arm geometry, sampling at 48 kHz and

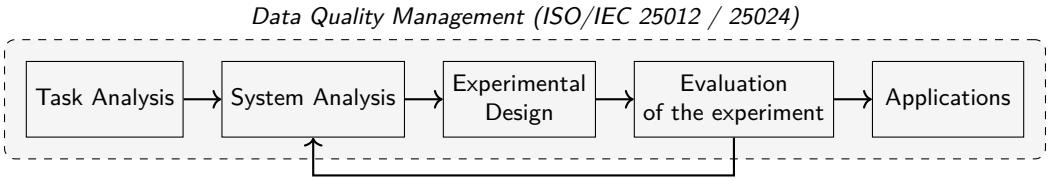


FIG 1. Workflow of the DoE process for identifying the causal relation between input factors and response variables. Data Quality Management (DQMa) acts as a cross-cutting layer: it constrains the design before execution and verifies the recorded data afterwards.

recording the UAV top-down. The core of the approach is acoustic beamforming, which spatially localizes sound sources: a functional beamformer operating in the frequency domain via a Cross Spectral Matrix for the Convolutional Neural Network (CNN) pipeline, and a rotating beamformer that isolates individual blade tips for the binary classifier (25). Processing runs in Acoular (Python), analyzing third-octave bands between roughly 6.3 and 12.5 kHz, with the Point Spread Function (PSF) used to manage resolution and artifacts.

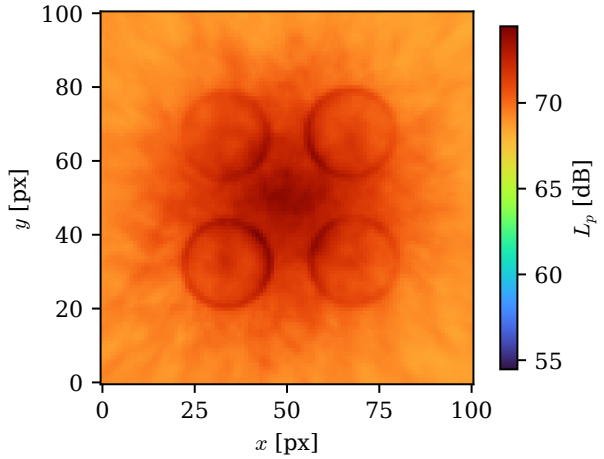


FIG 2. Acoustic image map for the quadrocopter with an intact propeller set-up at a beam forming frequency of 12.5kHz indicating the sound pressure level L_p. The motor-to-motor distance is 480mm and the length of the propellers is 250mm.

The beamforming output as shown in figure 2 takes the form of acoustic images — spatial source maps that serve as features for classification. Two paths follow: a CNN (TensorFlow/Keras) that classifies these images directly, and a simpler linear classifier that thresholds the peak ratio between the two blade-tip sources, requiring no machine learning. Taken together, the technology enables non-invasive, ground-based condition-based maintenance (CBM) forUAVs, designed for integration into a vertiport where propeller health can be checked automatically during landing or approach.

(25) investigates four third-octave bands, defined by their center frequencies of 6.300, 8.000, 10.000, and 12.500 Hz (corresponding to frequency ranges of 5.613–7.072, 7.127–8.980, 8.909–11.225, and 11.136–14.031 Hz). These higher bands were chosen because the broadband component of propeller noise — driven by blade-tip turbulence and therefore carrying the damage information — dominates there, while the smaller Beam Width (BW) allows sharper source localization. In the evaluation, the two higher bands (10.000 and 12.500 Hz) consistently perform best for CNN classification; the linear classifier's ω-estimation algorithm likewise operates in the 12.500 Hz band.

2.4.2. Data Quality Requirements for Acoustic Fault Detection

The results of steps 1) and 2) from Section 2.1.1 are presented here for our use-case. The data entities identified in step 1) are:

(i) an individual microphone sample,
(ii) the corresponding array sample,
(iii) the beamforming output image generated from the array sample,
(iv) a complete dataset consisting of many selected measured or synthesized beamforming output images (used as training or validation data for the ML process).

For these entities, the corresponding metadata and the synthesized equivalents must also be taken into account. To illustrate step 2), we specify as an example the two most common specific characteristics for these entities: completeness and accuracy. We list only those entities to which the characteristics apply.

Completeness Regarding the specified data entities, completeness means, for (i) no data loss occurs in an individual microphone sample, for (ii) all microphones provide usable samples, and for (iv) the existence of all relevant parameters and dimensions. Similarly, for the associated metadata, completeness means that all relevant attributes are stored and that there is no data loss or corruption.

Optimizing the completeness of a dataset in this sense is one of the main objectives of DoE. Other characteristics from the ISO 5259-2 standard that are relevant to DoE but are not discussed here are "balance" and "similarity".

Accuracy Concerning the specified data entities, accuracy means, for (i) the sample represents accurately the relevant acoustic information. In the case of synthesized data, the accuracy is concerned with the accuracy of the simulation. Analogously, for the corresponding metadata, accuracy means that all attributes are correct. This means, for example, that all relevant descriptions and the labelling of the measurement object are correct.

The only data entity considered in the remainder of this paper is "microphone array completeness" — case (ii). This was chosen as the first example to test the con-

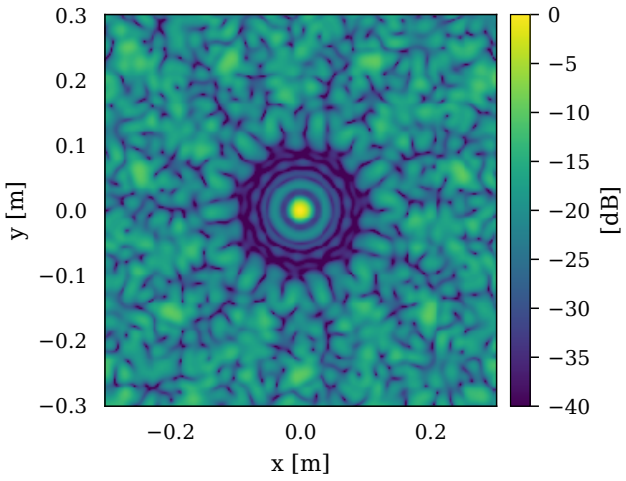


FIG 3. Analytical point spread function of the intact array at the analysed centre frequencies 12.500Hz, normalised to the peak level. BW = 1.5cm and DR = 9.72 dB

cepts of our proposed framework, since the metric for completeness is trivial and the data pollution can also be easily implemented. As measure of completeness, we use the ratio of usable individual microphone samples to the total number of microphones in the array. The pollution is generated by just retaining some microphone samples from the beamforming process. Both from the perspective of certification and in terms of actual use and the inference process, this use-case refers to situations in which data loss occurs in the detection system.

2.4.3. Beamforming and Image Quality Metrics

The beamforming map is the input representation the classifier actually sees, so its quality bounds what any downstream model can achieve. Assessing that quality requires metrics that respond to the perturbations expected in operation — above all a thinning of the array through failed or omitted channels. Table 2 lists the metrics used here.

The first three characterize the map itself. BW measures the achievable resolution and follows directly from aperture and frequency. It states how far apart two sources must be to remain separable. Dynamic Range (DR) captures the level difference between the main lobe and the side lobes, and is the metric most sensitive to a thinned array: removing microphones lifts the side lobes, compresses the DR and introduces artefacts that are indistinguishable from genuine sources. The third metric addresses a different question. The correlation of the PSF to a reference PSF, expressed as a Pearson coefficient, quantifies the overall change in imaging geometry rather than a single property of the map. It matters because a CNN trained on one PSF does not adapt to an altered one. The correlation therefore serves as a proxy for model degradation and links map quality to detection performance.

This distinction is the reason the three metrics are not interchangeable. BW, DR and mean are properties of the representation and can be evaluated without any model.

TAB 2. Performance metrics for the data assessment.

Metric	Function / Impact
Beamwidth (BW)	Resolution; depends on aperture and frequency
Dynamic range (DR)	Side-lobe artefacts; a thinned array lifts the side lobes, lowering dynamic range and introducing further artefacts
PSF correlation (r_{PSF})	Change in imaging geometry, evaluated on the simulated point spread function of a single reference source
Acoustic map correlation (r_{AIM})	Same measure applied to the measured acoustic map of the test object (25)

Only the PSF correlation has so far been connected to a measured loss in classification performance. Used as a data quality criterion, the three are therefore necessary conditions with no established threshold, whereas the PSF correlation is the only one for which a target value can be argued from an accuracy requirement.

The PSF is formulated as:

$$h_m(\mathbf{x}) = \frac{e^{-jk\|\mathbf{x}-\mathbf{r}_m\|}}{4\pi\|\mathbf{x}-\mathbf{r}_m\|}, \tag{1}$$

where $\mathbf{x}$ denotes the focus point, $\mathbf{r}_m$ the position of the m-th microphone, and $k = \omega/c$ the wavenumber.

A single monopole sits at focus. A conventional frequency-domain beamformer — the $\gamma = 1$ limit of the Functional Beamformer used for the measured maps — is steered across the mapping plane at the same, resulting in figure 3 In order to compare further any impact on the performance, the following performance metrices are derived and listed in table 2.

3. RESULTS

This study rests on following deliberate scope decisions: First, the analysis is limited to a single frequency band at 12.500Hz. the higher frequencies have proven to be more successful during ML-training (25). The study quantifies how sensitive the acoustic images are to missing sensor information, thereby characterising the robustness of the measurement chain under data constraints. Second, the fault mode for data is considered to be a cut off propeller blade. However, in the following part, a complete healthy set-up is used.

I. - Sensitivity Analysis at Inherent Level

The objective is to examine the sensitivity of the tool towards incomplete sensor data. Each data set represents the respective missing sensor and its impact on the overall results. For this the average loudness is measured and compared against the benchmark which equals all sensors in progress. For each dropout count, channels are removed and two metrics are evaluated: the DR (level difference between the main lobe and the highest sidelobe) illustrated in figure 4 for DR and in figure 5 for PSF. The

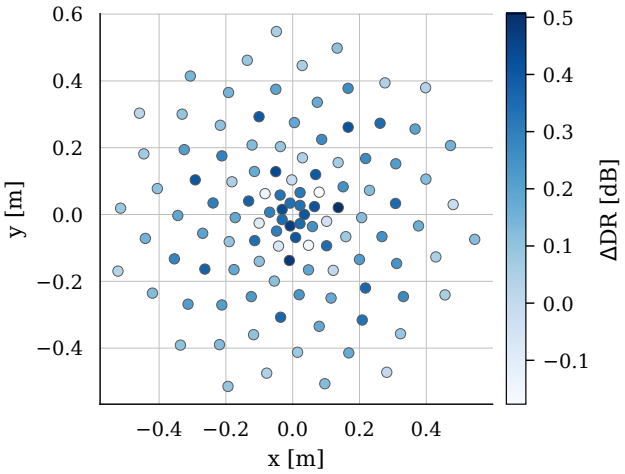


FIG 4. Contribution of each microphone to the dynamic range, obtained by removing one channel at a time from the intact 112-element array.

PSF correlation defines the correlation between the point spread functions of the degraded and the intact array as presented in table 2. Besides the spatial impact analysis for a single failure, an increasing sensor performance is investigated and illustrated in figure 6. Both are computed over increasing dropout realisations, giving a mean and a standard deviation (shaded bands). Two tolerance criteria — a dynamic-range reduction of at most 2 dB and a PSF correlation of at least 0.95 — are imposed, and the dropout count at which each is first violated is obtained by interpolation (vertical lines). The analysis is performed at the PSF level, so the resulting limits characterise an intrinsic property of the array rather than a specific measurement.

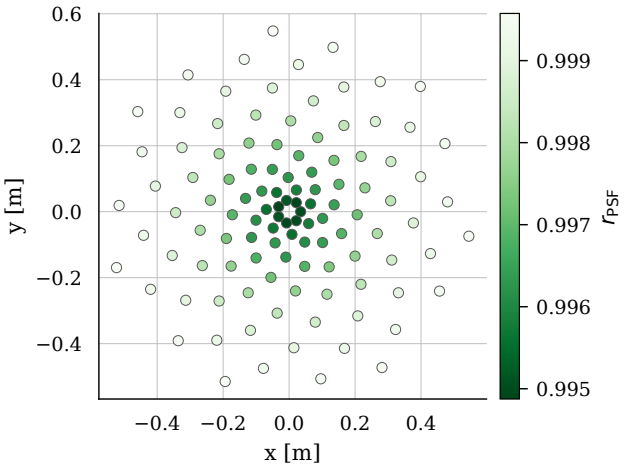
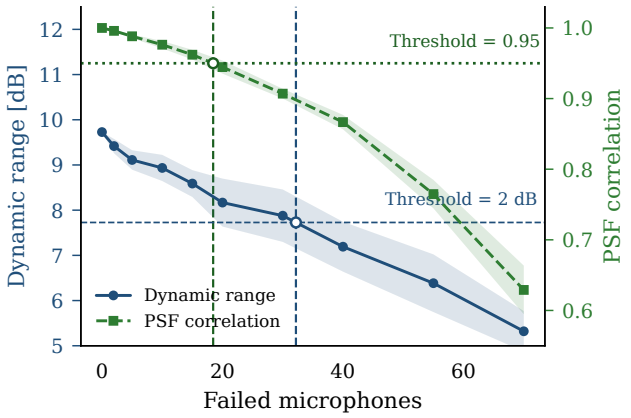


FIG 5. Contribution of each microphone to the point spread function, obtained by removing one channel at a time from the intact 112-element array.

The comparison at a fixed count of 20 failed channels distinguishes five spatial patterns of failure, chosen to span the ways in which channels are lost in practice rather than to sample uniformly. *Random* removes 20 channels drawn uniformly from the 112 elements without regard to position and serves as the reference case used in the preceding sweep *Arm-wise* removes complete arms of the spiral layout, corresponding to the loss of a wiring harness



(a) PSF correlation

FIG 6. Dynamic range (left axis) and PSF correlation (right axis) as a function of the number of randomly failed microphone channels. The markers represent the indicated metric over a decreasing number of failed microphones. The shaded bands represent one standard deviation. Horizontal dashed lines mark the tolerance criteria (-2 dB relative to the intact array, and a PSF correlation of 0.95) whereas the vertical lines mark the interpolated crossing of each criterion.

or of one segment of the array frame. *Outer* removes the 20 elements at the largest radii, *inner* the 20 elements closest to the array centre; these two isolate the roles of the radial position, since the outer elements determine the effective aperture while the inner elements populate the short baselines of the co-array. *One-sided* removes all 20 channels from one half-plane of the array, representing a localized mechanical damage or a partial mounting failure. The results in figure 7 show that the two metrics respond differently to the pattern. Dynamic range varies only moderately, between 7.9 and 9.72 dB so the level relation between main and side lobes is governed mainly by the number of missing channels rather than by their arrangement. The PSF correlation, in contrast, spans 0.87 to 0.975 across the same configurations and is therefore the pattern-sensitive criterion.

The ordering is worth stating explicitly, because it does not follow the intuition that clustered failures are uniformly more severe than scattered ones. One-sided (0.963) and outer (0.975) losses preserve the shape of the point spread function better than random loss (0.940), whereas arm-wise (0.928) and above all inner failures (0.87) fall clearly below it. The inner elements sample the short baselines that suppress the side-lobe structure; removing them alters the shape of the map even though the aperture, and with it the beamwidth, is retained. What matters for the tolerance criterion is thus not clustering as such but which part of the aperture is lost. Any acceptance criterion applied at run-time therefore has to consider the position of the failed channels, not only their count.

The diagnostic model operates on acoustic maps such as those shown in Figure 2, and the sensitivity study is therefore repeated on these maps, using the acoustic map correlation r_{AIM} against the fully functional array as reference. Where the point spread function isolates

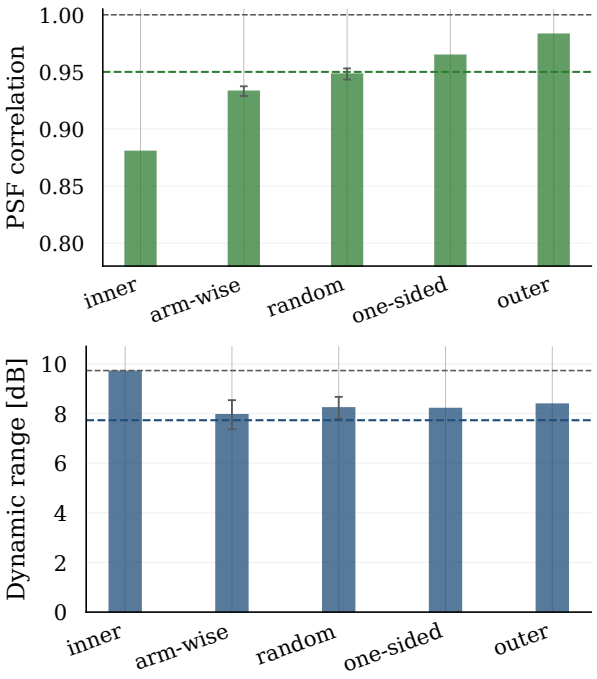


FIG 7. Effect of the spatial failure pattern for 20 channels, ordered by PSF correlation: Dashed lines mark the intact array and the respective tolerance criteria

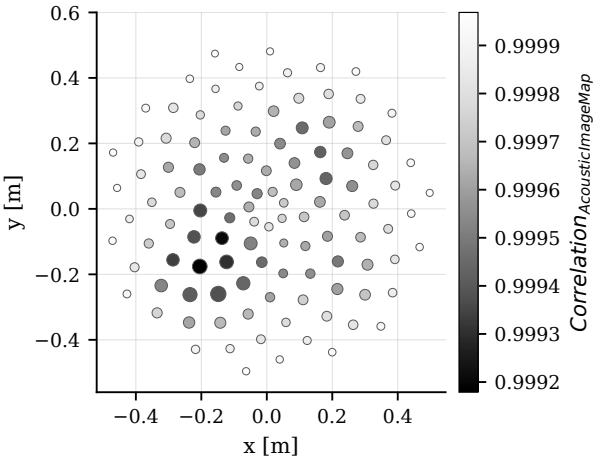


FIG 8. Dropout sensitivity per microphone position regarding their impact on the acoustic image map, combining effect magnitude and uncertainty. Colour encodes the correlation between baseline and modified map; marker size encodes the uncertainty of the estimate (standard deviation across comparison pairs). Frequency band: 12.5 kHz.

the imaging geometry of an idealized single source, the acoustic map exposes the same degradation to a real, extended source distribution — and thus reveals how much of a channel loss the beamforming itself absorbs. The underlying recordings were taken in an anechoic chamber and capture the complete UAV rather than the propeller alone; what the maps show is accordingly the system dynamics in which the fault mode is embedded.

The evaluation proceeds per microphone position. For each single-channel dropout, the correlation between the baseline and the modified map is averaged over all matched comparison pairs and assigned to the affected channel, yielding a spatial sensitivity map across the array Figure 8. The result is displayed on the physical array geometry, with colour encoding the metric at each position, while a second representation adds the uncertainty of the estimate — the standard deviation across pairs — through marker size, so that magnitude and variability can be read from a single figure.

4. ANALYSIS

Two tolerance criteria are applied to the thinned array. A dynamic-range loss of at most 2.0,dB is maintained up to 34 failed channels, whereas the PSF correlation criterion of at least 0.95 is already violated at 20 failed channels. The PSF correlation is therefore the binding criterion.

This limit holds under the following restrictions: It applies to randomly distributed failures. Clustered or inner losses remove a region of the aperture and are more severe, as the pattern analysis in figure 7 confirms. It was obtained at a single frequency, and since the array pattern scales with frequency relative to the aperture, the sweep must be repeated at the lowest and highest band before the limit can be quoted for the full configuration. Finally, both criteria describe the beamforming map rather than the diagnosis performance: a network trained on the intact array is fitted to the specific point spread function. Consequently, a robust classification performance strongly sensitive to a loss.

The evaluation proceeds in two steps, which answer different questions and are therefore reported separately. The first is a sweep over the number of failed channels. Starting from the intact 112-element array, channels are removed at random positions in increasing numbers, the beamforming map is recomputed for each configuration, and the metrics introduced in section 3 are evaluated against the intact array as reference. Repeating the draw at each failure count separates the systematic trend from the variability of the particular set of removed channels. This step establishes which metric is the binding criterion and yields a first tolerance limit. The second step fixes the number of failures at 20, slightly beyond the armwise-failure limit so that the spatial effect is clearly visible, and varies the pattern in which they occur: inner, arm-wise, random, outer and one-sided. Its purpose is to test whether the limit obtained from random failures transfers to the failure modes that actually occur in operation, where channels are rarely lost independently of one another.

Both steps evaluate properties of the beamforming map rather than of the diagnosis itself. They bound the region in which the input representation remains comparable to the one the model was trained on; whether classification performance is retained within that region is a separate question.

5. DISCUSSION AND OUTLOOK

The results permit two kinds of statement, and it is worth keeping them apart. The sweeps establish how far the beamforming map may be degraded before it leaves the envelope in which it resembles the data the model was

trained on. They do not establish how far the diagnosis may be degraded, and the gap between these two questions frames what follows.

The tolerance limits translate directly into an input-quality gate. Channel loss is detectable at run time against a validity criterion, and the limit derived here 20 of 112 channels for randomly distributed failures can be evaluated before a measurement is passed to the classifier. This is the practically useful outcome: a criterion that requires no model inference and no labelled data, and that can therefore be applied in operation. The pattern analysis qualifies this. The admissible number of failures depends on which part of the aperture is lost, and the assumption that clustered failures are uniformly more severe than scattered ones does not hold. One-sided and outer losses preserve the imaging geometry better than random loss, whereas the loss of inner elements those populating the short baselines that suppress the side-lobe structure is substantially more damaging. An acceptance criterion applied at run time must therefore consider the positions of the failed channels, not only their number.

A third implication concerns the design of the array itself. The leave-one-out analysis locates the contribution of each element, and the resulting ranking identifies elements whose loss is disproportionately harmful. This information is available before deployment and can inform redundancy decisions — which channels warrant duplicated wiring, which are candidates for a reduced-cost variant.

Three limitations bound the results, and one of them is a knowledge gap rather than a methodological choice.

The first is the one already stated: all criteria are map-level. A network fitted to the point spread function of the intact array has no mechanism to compensate a change in it, so classification performance may degrade before any map-level threshold is violated. Establishing the operationally relevant limit requires the same sweep evaluated on the model output. Approaches for quantifying the generalization gap of a neural network offer one route to this.

The second limitation is that the fault signatures themselves are only partially characterized. The response space was reduced to two physical mechanisms that can be varied independently in the experimental design — distortion induced by forward flight, and distortion induced by the installation, in particular rotor–strut interaction. This reduction is deliberate and makes the design tractable, but it presupposes knowledge that is incomplete: how bearing wear manifests acoustically, how the harmonic structure changes under progressive motor degradation, and what distinguishes the signature of blade damage from an altered operating point. Until these are established experimentally, the factor space of the design is defined by what can be varied rather than by what is diagnostically relevant — and that is precisely the assumption that the framework in Section 2.1 identified as testable but not guaranteed.

The third limitation is the restriction to a single frequency and to the specific array geometry. Both the tolerance limits and the element ranking are properties of this configuration and do not transfer without re-evaluation.

6. CONCLUSION

This work proposed a three-level framework for assessing dataset quality in acoustic fault detection, intrinsic properties, experimental design, and fitness for the learning task, which differ in when they can be verified. DoE therefore does not guarantee high-quality data; it maximizes the expected information per run, provided the modelled factor space contains the space encountered in operation.

Applied to the sensitivity of beamforming maps to microphone failure, the PSF correlation proved the binding criterion, yielding a limit of 20 of 112 channels for random failures. The limit is pattern-dependent: the loss of inner elements is considerably more damaging than one-sided or outer loss. Since both criteria describe the map rather than the diagnosis, the limit is a screening threshold for input quality, not a validated bound on diagnostic performance.

The broader contribution is the direction of effect: DoE is the only instrument by which data quality can be influenced before acquisition, while the deviations catalogued by data quality management are candidate factors for the design itself.

6.1. Outlook

A DQMo establishing measures, target value and evidence for each data entity remains to be completed. The more interesting question follows from it: which quality characteristics act as constraints on experimental design rather than as post-hoc acceptance criteria. Identifying these characteristics would couple the two disciplines, instead of merely aligning them.

Experimentally, the analysis was carried out at a single frequency; a multi-band treatment and a full DoE over the identified factor space remain open. This leaves open the question that this work was intended to address — whether systematically designed data generation yields information that data volume alone does not; this requires a comparison with a non-designed campaign of equal cost.

Declaration of Generative AI in Writing Process

During the preparation of this work the authors used a large language model to improve the readability and language of individual passages. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] Christopher Teubert, Ahmad Ali Pohya, and George Gorospe. An analysis of barriers preventing the widespread adoption of predictive and prescriptive maintenance in aviation (1). 2023.

[2] Joseph M. Juran. *Juran's Quality Handbook*. McGraw-Hill, 5th edition, 1999. ISBN: 0-07-034003-X.

[3] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, (12):5–33, 1996.

[4] Thomas C. Redman. *Data quality: The field guide*. Digital Press, Boston, 2001. ISBN: 9781555582517.

[5] N. Omri, Z. Al Masry, N. Mairot, S. Giampiccolo, and N. Zerhouni. Towards an adapted phm approach: Data quality requirements methodology for fault detection applications. *Computers in Industry*, 127:103414, 2021. DOI: 10.1016/j.compind.2021.103414.

[6] Mariana Salinas-Camus, Kai Goebel, and Nick Eleftheroglou. A comprehensive review and evaluation framework for data-driven prognostics: Uncertainty, robustness, interpretability, and feasibility. *Mechanical Systems and Signal Processing*, 237:113015, 2025. DOI: 10.1016/j.ymssp.2025.113015.

[7] EASA. NPA 2025-07 (B) — Proposed detailed specifications and AMC & GM for AI trustworthiness (DS.AI).

[8] Rosa Arboretti, Riccardo Ceccato, Luca Pegoraro, and Luigi Salmaso. Design of experiments and machine learning for product innovation: A systematic literature review. *Quality and Reliability Engineering International*, 38(2):1131–1156, 2022. DOI: 10.1002/qre.3025.

[9] Johannes Freiesleben, Jan Keim, and Markus Grutsch. Machine learning and design of experiments: Alternative approaches or complementary methodologies for quality improvement? *Quality and Reliability Engineering International*, 36(6):1837–1848, 2020. DOI: 10.1002/qre.2579.

[10] Douglas C. Montgomery. *Design and analysis of experiments*. Wiley, Hoboken, NJ, tenth edition, emea edition edition, 2020. ISBN: 9781119816959.

[11] Rosa Arboretti, Riccardo Ceccato, Luca Pegoraro, and Luigi Salmaso. Design of experiments and machine learning for product innovation: A systematic literature review. *Quality and Reliability Engineering International*, 38(2):1131–1156, 2022. DOI: 10.1002/qre.3025.

[12] Rosa Arboretti, Riccardo Ceccato, Luca Pegoraro, and Luigi Salmaso. Design choice and machine learning model performances. *Quality and Reliability Engineering International*, 38(7):3357–3378, 2022. DOI: 10.1002/qre.3123.

[13] ISO/IEC 25012:2008. Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model.

[14] Andrew Black, Peter van Nederpelt. Dimensions of Data Quality (DDQ): Research Paper, 2020.

[15] ISO/IEC 5259-2:2024. Artificial intelligence - Data quality for analytics and machine learning (ML): Part 2: Data quality measures.

[16] Peter van Nederpelt, Andrew Black. Dictionary of dimensions of data quality (3DQ): Dictionary of 60 Standardized Definitions. 2020.

[17] Leo Pipino, Richard Wang, David Kopcso, William Rybolt. Developing measurement scales for data-quality dimensions. In Y. Richard Wang, editor, *Information quality*, Advances in management information systems. Routledge, Abingdon, Oxon, 2015.

[18] ISO/IEC 25024:2015. Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality.

[19] ISO/IEC 5259-2:2024. Data quality: Part 210: Sensor data: Data quality characteristics.

[20] ISO/IEC 5259-2:2024. Data quality: Part 220: Sensor data: Quality measurement.

[21] ISO 5305:2024. Noise measurement for UAS (unmanned aircraft systems).

[22] RTCA. Standards for processing aeronautical data. Technical Report DO-200C, RTCA, Inc., Washington, D.C., 2024.

[23] Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. The effects of data quality on machine learning performance.

[24] Anonymous. 11. design of experiments (doe): Quality management in theboschgroup |technical statistics: Bve 1724-1, 10.2010.

[25] Leon Steinhoff, Ann-Kathrin Koschlik, Emy Arts, Maria Soria-Gomez, Florian Raddatz, and Veit Dominik Kunz. Development of an acoustic fault diagnosis system for uav propeller blades. *CEAS Aeronautical Journal*, 15(4):881–893, 2024. DOI: 10.1007/s13272-024-00752-8.